hirid data - observation data

HIRID Variables - SOFA calculation

- SOFA 所需的变量全部可得到，具体可参考 Sepsis 3 Definition

- pao2fio2ratio 和 GCS 两个变量需通过计算间接得到

HIRID Variabes - PM_sepsis_shock_prediction

- **sepsis_shock_prediction** : age, gender, temperature, spo2, sbp, dbp, mbp, heart_rate, resp_rate, lactate, aniongap,wbc, bicarbonate, gcs_motor, gcs_verbal, gcs_eyes, inr, pt, ptt, baseexcess, po2, creatinine, potassium, ph, bun, glucose, totalco2, platelet, hematocrit, gcs, rbc, calcium, urineoutput

- **PM_sepsis_shock_prediction有hirid无**：aniongap, totalco2, hematocrit(红细胞比容)，rbc，gcs（通过计算可间接得到）

Raw HIRID data processing

（1）将250个大csv文件划分为以指标命名的小csv文件

`python $bin/step01_divideFile_by_var2.py /home/hanl/sepsis/00.data/HIRID_data/hirid_pre/observation_tables/csv/ $wd/var/`

▼ Input:

  (1) 需要切分csv所在的目录

  (2) 输出目录：xxxx/var/

▼ Output

☐ step02： 根据MIMIC所需的表转为字典格式即{表名1: [指标1，指标2.......], 表名2:[指标1，指标2.....]......}；表名形成目录，每个目录下存在一个 <表名_merge.txt>, 里面包括该表对应指标在step01输出目录中的所有路径

```
python $bin/step02_VarTabCombine.py $wd/input/varidx_varname.csv $wd/var $wd/tab
```

▼ Input

（1）csv文件（两列）：varname，varid （存在一个name对应多个idx的现象）

```
carboxyhemoglobin,24000526
chloride,24000439
chloride,24000521
creatinine_in_urine,24000572
creatinine_in_urine,24000573
creatinine,20000600
dbp,120
dbp_ni,620
```

（2）step01输出路径

（3）输出目录：：xxxx/tab/

（4）MIMIC所需的表的表头转成的字典格式: /input/mimi4_var.dict

▼ output: 如下

```
$ll tab/*/*merge.txt
-rw-rw-r-- 1 hanl hanl 629680 Jan 29 17:48 tab/bg/bg_merge.txt
-rw-rw-r-- 1 hanl hanl  68560 Jan 29 17:48 tab/blood_differential/blood_differential_merge.txt
-rw-rw-r-- 1 hanl hanl 359940 Jan 29 17:48 tab/chemistry/chemistry_merge.txt
-rw-rw-r-- 1 hanl hanl  68560 Jan 29 17:48 tab/coagulation/coagulation_merge.txt
-rw-rw-r-- 1 hanl hanl 137120 Jan 29 17:48 tab/complete_blood_count/complete_blood_count_merge.txt
-rw-rw-r-- 1 hanl hanl  51420 Jan 29 17:48 tab/gcs/gcs_merge.txt
-rw-rw-r-- 1 hanl hanl  17140 Jan 29 17:48 tab/height/height_merge.txt
-rw-rw-r-- 1 hanl hanl  17140 Jan 29 17:48 tab/urine_output/urine_output_merge.txt
-rw-rw-r-- 1 hanl hanl 391360 Jan 29 17:48 tab/vitalsign/vitalsign_merge.txt
```

```
/home/hanl/sepsis/00.data/HIRID_data/hirid/var/part-52/24000572.csv
/home/hanl/sepsis/00.data/HIRID_data/hirid/var/part-245/24000572.csv
/home/hanl/sepsis/00.data/HIRID_data/hirid/var/part-202/24000572.csv
tab/chemistry/chemistry_merge.txt
```

☐ step03

```
python $bin/step03_TabConvert.py $wd/tab/
```

☐

1) INPUT 目录中的文件们

2) workshell

```
wd=/home/hanl/sepsis/00.data/HIRID_data/hirid/
bin=/home/hanl/sepsis/00.data/HIRID_data/bin/

###pre
#mkdir var tab

# file : varname varid
python $bin/get_var_NameID.py $wd/input/HIRID_var2idx.csv
# file : MIMIC4 tab var_list
/home/hanl/software/Python-3.7.4/bin/python3.7 $bin/get_FinalCsvHeader.py
python $bin/obser_var_stat.py $wd/input/MIMIC_output_demo/
python $bin/get_idName.py $wd/input/varidx_varname.csv

### deal
python $bin/step01_divideFile_by_var2.py /home/hanl/sepsis/00.data/HIRID_data/hirid_pre/observation_tables/csv/ $wd/var/
python $bin/step02_VarTabCombine.py $wd/input/varidx_varname.csv $wd/var/ $wd/tab/
python $bin/step03_TabConvert.py $wd/tab/
python $bin/step04_final_out_stat.py $wd/tab
```

Raw data

T1 - observation_tables_index

- 两列变量都需要

T2 - general_table

	patientid	admissiontime	sex	age	discharge_status
2	1	2197/4/3 19:15	M	75	alive
3	2	2117/10/18 22:35	M	80	alive
4	3	2163/10/17 19:15	M	55	alive
5	4	2149/1/8 15:30	M	75	alive
6	5	2144/6/6 16:15	M	20	alive
7	6	2143/9/4 15:15	F	65	alive
8	7	2116/11/22 18:15	M	65	alive
9	8	2139/12/21 12:40	M	60	alive

general_table

	Source Table	variableid	Variable Name	变量名称	Unit
1	Source Table	variableid	Variable Name	变量名称	Unit
2	Observation	200	Heart rate	心率	/min
3	Observation	410	Core body temperature	核心体温	°C
4	Observation	7100	Rectal temperature	直肠温度	°C
5	Observation	400	Axillary temperature	腋下温度	°C
6	Observation	100	Invasive systolic arterial pressure	有创收缩压	mmHg
7	Observation	120	Invasive diastolic arterial pressure	有创舒张压	mmHg
8	Observation	110	Invasive mean arterial pressure	有创平均动脉	mmHg
9	Observation	600	Non-invasive systolic arterial pressure	无创收缩动脉	mmHg
10	Observation	620	Non-invasive diastolic arterial pressure	无创舒张压	mmHg
11	Observation	610	Non-invasive mean arterial pressure	无创平均动脉	mmHg

hirid_variable_reference

	A	B	C	D	E	F	G	H	
1	datetime	entertime	patientid	status	stringval	type	value	variableid	
2	00:00.0	20:27.0	148	8			0	30005080	
3	00:00.0	16:11.6	148	8			0	30005110	
4	00:00.0	20:26.9	148	8			0	30005010	
5	00:00.0	20:27.0	148	8			0	30005075	
6	35:00.0	36:37.7	148	12			73	4000	
7	35:00.0	36:37.6	148	12			98	200	
8	36:00.0	37:37.7	148	8			94	4000	
9	36:00.0	37:37.7	148	8			118	200	
10	37:00.0	38:37.2	148	12			-2.5	211	
11	37:00.0	38:37.2	148	8			94	4000	

part-0

数据整理

- 将目标MIMIC4中需要提取的表及其中的指标输出为字典

```
{vitalsign:['temperature', 'dbp_ni', 'sbp', 'temperature_site', 'spo2', 'sbp_ni', 'glucose', 'heart_rate', 'dbp', 'mbp_ni', 'mbp', 'resp_rate']
coagulation:['thrombin', 'fibrinogen', 'inr', 'pt', 'ptt', 'd_dimer']
crrt:['clots', 'blood_flow', 'dialysate_fluid', 'heparin_concentration', 'prefilter_replacement_rate', 'replacement_fluid', 'ultrafiltrate_outpu
urine_output:['urineoutput']
chemistry:['albumin', 'chloride', 'calcium', 'glucose', 'bun', 'creatinine', 'potassium', 'total_protein', 'aniongap', 'globulin', 'sodium', 'bi
gcs:['gcs', 'gcs_verbal', 'gcs_motor', 'gcs_unable', 'gcs_eyes']
height:['height']
bg:['aado2_calc', 'ph', 'methemoglobin', 'po2', 'hemoglobin', 'pco2', 'baseexcess', 'potassium', 'fio2_chartevents', 'specimen', 'carboxyhemoglo
blood_differential:['bands', 'monocytes', 'atypical_lymphocytes', 'nrbc', 'lymphocytes', 'eosinophils', 'eosinophils_abs', 'neutrophils_abs', 'i
complete_blood_count:['mcv', 'rdw', 'hemoglobin', 'platelet', 'mchc', 'rdwsd', 'hematocrit', 'mch', 'rbc', 'wbc']
}
```

-